Transcription Rules

Metadata (internal)

- Each speaker should be identified by name and last name
- Each speaker should have and ID

Quality of sound: (out of 5)

```
- 1 Undistinguishable
- 2 Major levels of saturation/background noise/distinguishable with
difficulty (10% of UNK workds at most).
- 3 High levels of saturation/background noise/distinguishable with some
difficulty (5% of UNK workds at most).
- 4 Low level of saturation/distinguishable with no efforts (less than 1% of
UNK workds at most).
- 5 High quality audio.
```

Sample of Metadata

```
; Date of speech:   01/15/2013
; Location of Speech:   QNNC, Doha, Qatar
; Occasion of Speech:   Arabic NLP Conference
; Keywords: Education, Qatar, Doha, Strategy
; Sound Quality:    4/5
; Last Update Date: 01/25/2013
; Transcriber:      XYZ
; Comments:     Speaker ABC stutters
; Comments:     Disturbances between turn 004 and 020
```

Segments

- No more than 1 second of silence in a segment
- Maximum of 15 seconds of length
- Maximum of 40 characters per line
- No more than 80 characters per segment
- No more than 2 lines of text in a segment

Segment Correctness

Segments must not contain any kind of errors (typos, spelling errors).

Segment Coherence

Longer segments should be divided into smaller units containing whole words and assuring a minimum of syntactic and semantic coherence

[Example from TED]

INCORRECT:

```
S1: somehow, this worked really well
    in her garage. When you work
S2: on something big,
    you need to accept failure.
```

CORRECT:

```
S1: somehow, this worked really well
    in her garage.
S2: when you work on something big,
    you need to  accept  failure.
```

Transcription rules

1) Write every speech event. Including hesitations, repetitions and onomatopoeias.

```
S1: [Umm] he said that that was fine
```

2) Write the words in the appropriate language with the correct orthographical form (avoid typos and spelling mistakes). Foreign words and phrases need to be annotated (see transcribing foreign words)

```
S1: She started singing [FOR:FR la vie en rose]
```

3) Do not transcribe contractions

```
WRONG:  I'm ready
CORRECT: I am ready
```

4) Transcribe in a consistent and homogeneous manner. Proper and common nouns should be transcribed consistently throughout the project.

```
WRONG: S1:   نا من مدينة شيكاغو,
 وتقع شيكاجو في شمال امريكا
```

5) Do not use Capitalized letters at the beginning of the segments.

```
WRONG: She is going to the show
CORRECT: she is going to the show
```

6) Reserve capitalization for acronyms, spellings and proper nouns

```
she is going to Wisconsin, in the USA
```

7) Acronyms are words using only letters as abbreviations for longer expressions

```
he just bought a new CD
```

Punctuation

- Only use "." "," "?" ":" ";" "!" "…" and double-quotes (" ") either in English or Arabic
- Do not split compounds (E.g. middle-aged, five-sided, African-American)
- Do not tokenize the apostrophe. Eg. Write Gordon's instead of Gordon 's.

Foreign words

- Transcription of foreign words that do not use the same script system should be transliterated, and such transliteration should be consistent throughout the document/video.

List of symbols

@ are the words

Speech - Dialogue

1 **False start** | [FALSE @]

> *The act of beginning an utterance but aborting before completion*

```
well,  [FALSE I was trying ] I managed to get here just in time

[FALSE what do you] yeah I mean the most ...
```

2 Repetition or correction | [REP @]

> *When the speaker stutters or Re-expresses things during moments of excitement or uncertainty and subsequently correct themselves.*

```
if I [REP can't find] don't know the answer myself, I will find it
[REP if] if
[REP it was a] [REP it was] I think it was a [NE Honda]
```

3 Interruption | [INTERP @[X]@]

> *Interruption of one or several lexical units caused by any kind of disruption (technical or human related)*

These will be annotated in two ways:

- When the lexical item is complete, or recoverable:

  ```
  This is un[NOISE]beliv[BREATHE]able
  ```

- When there is a missing part or it is incomplete

  ```
  This is [INTER un[X]eable]
  You are [INTER wha[X]]
  ```

4 Hesitation| [HES]

> *Any kind of sound produced during spontaneous speech that represent a pause filled by vocalization. (Excamples uh, uhm, mmm,)*

```
yes [HES] yes
but [HES] I don't know if you want to ask to [HES] each person
```

5 Interjection | [INTERJ]

> *Expression of surprise (Oh), affirmation (uh-huh), or negation (uh-uh) ENGLISH: Positive: Voiced: uh-huh, Non-voiced: mm-mhm Negation: Voiced: uh-uh, Non-voiced: mm*

ARABIC:

[INTERJ uh-huh]. [HES] once we hit the record button we just let it go, no matter what [INTERJ uh-uh]. at least, I don't think so

Non Lexical Noises

> *These are non lexical acoustic events. Can be human like breathing, laugh or applause. Or can be non-human as music or any other undescribed noise.*

- human:

Breathing | [BREATH] |

Laugh | [LAUGH]

Applause | [APPLAUSE]

- non-human:

Noise | [NOISE]

Music | [MUSIC]

```
he was tired of [BREATH] keeping records.
[LAUGH] and the funny thing was that ...
```

Lexical notation

6 Foreign word | [FOR:LANG @]

> *Whenever a foreign word is used. The original language needs to be specified. Use the ISO 639-1 standard*

```
she started singing [FOR:FR la vie en rose], followed by [FOR:ES la bamba]
```

> *Transcription of foreign words and named entities that do not use the same script system should be transliterated, and such transliteration should be consistent throughout the entire document/video. Furthermore, a bank of transliterations should be shared and used across transcriptions.*

```
و قال [اجنبي:فر سي لا في]

هو يعمل في [ميكروسوفت علم:منظمة] في امريكا
```

< Also dialectal word in Arabic need to be annotated، شمي ، مصري، مغربي ، خليجي، أخرى

```
[لهجة:مصري كده]
```

8 Named Entities Proper Names | [NE:TYPE @]

> *Proper names are capitalized. Four types of named entities: Person (PER), location(LOC), organization (ORG), miscellaneous (MISC)*

```
[NE:PER Henry Ford] used to live in [NE:LOC Fifth Avenue] in [NE:LOC New
York] and created [NE:ORG Ford Motor Company]

[علم:شخص محمد عباس]  وسل [علم:مكان مطار حمد الدولي] في [علم:مكان الدوحة]
```

Non identifiable

Unidentifiable| [?UNK]

> *Reserve this tags for all those words that are uncertain and/or unidin*

Tag list

English Tag	Arabic Tag	Hot keys	Description
[FALSE @]	[خطأ: @]		False start
[REP @]	[إعادة:@]		Repitition or correction
[INTER @]	[توقف:@]		Interruption
[HES]	[تردد:@]		Hesitation
[INTERJ]	[تعجب:@]		Interjection
Non-Lexacal Noises			
[BREATH]	[تنفس]		Breathing
[LAUG]	[ضحك]		Laugh
[APPLAUSE]	[تصفيق]		Applause
[MUSIC]	[موسيقى]		Music
[NOISE]	[ضجيج]		Noise
[FOR:LANG @]	[أجنبي:لغة @]		Foreign word
[NE:PER @]	[علم:شخص @]		Named Entity Person
[NE:LOC @]	[علم:مكان @]		Named Entity Location
[NE:ORG @]	[علم:منظمة @]		Named Entity Organization
[NE:MISC @]	[علم:آخر @]		Named Entity Misclenious
[UNK]	[مبهم]		Unidentifiable